Capstone Project-2

Mechanical Fitting Failure Classification

Industry: Gas Pipeline

Prepared by: Prashant Sanghal

Why do this Project?

- Code of Federal Regulations (49 CFR Parts 191, 192) requires gas distribution pipeline operators to report hazardous leaks Involving mechanical fitting (DOT Form PHMSA F-7100.1-2).

- Oldest fitting installation was done back in 1851. Some 165 years ago.

So, using data can we find out:

What might have caused the leak?
When?
Where?
How often?
Was it the same fitting?
Same Manufacturer?
And ….so much more?

Benefit:

- Pipeline operators avoids environmental hazards.
- Manufacturers re-design improved fittings.
- Can set up early notification before leak occurs.
- Transport gas to various locations reliably.

What is our Data Story?

- Was available on Kaggle
- CSV format.
- 85,611 observation and 54 columns.
- All given in text.
- Many missing values (null, others, unavailable).
- Some critical information in other columns.

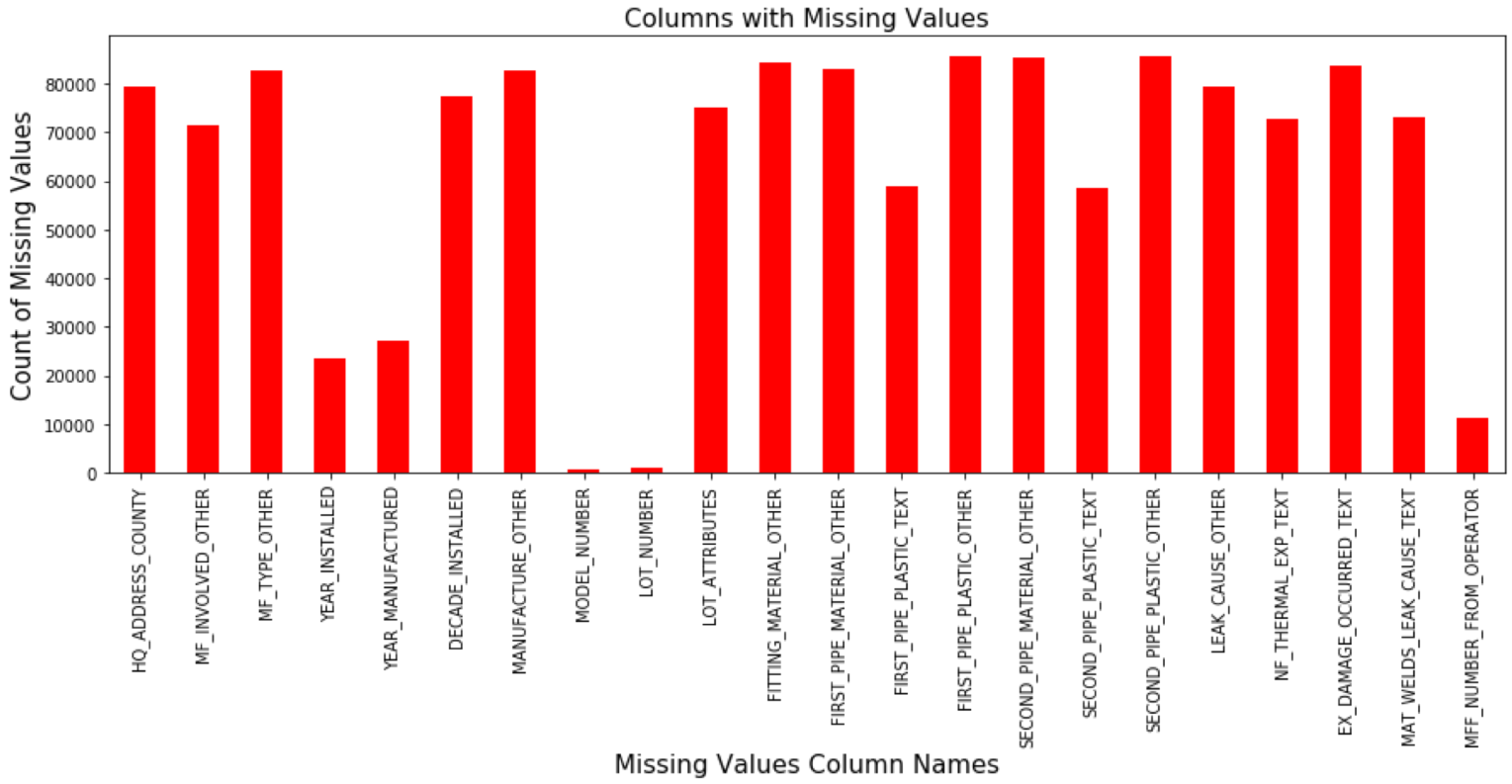
Question:

How to bring it all together?

- 15 out of 22 columns missing > 70% values
- Available data in missing columns have unique values
- Describing material Vs design defect, cause of leak
- Can we drop these columns?

Let’s build a Missing Dashboard:

- A function
- Returns summary table
- Shows missing values, missing percentage, unique values in missing columns.




```
Average (base) Missing Percentage: 0.73
Number of Missing Columns
(Random with Missing Percentage): 15
```

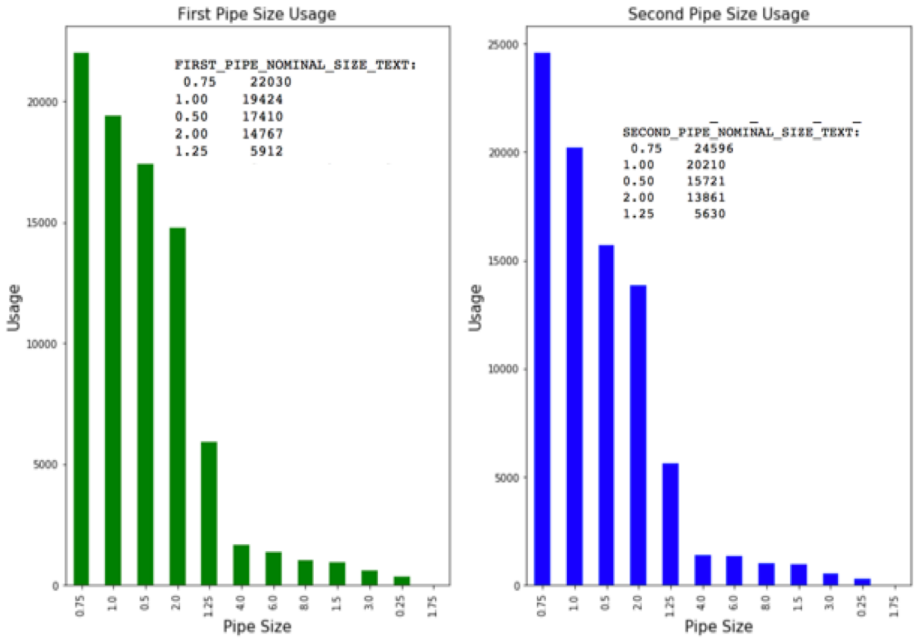
	missing_values	missing_percentage	unique_values_in_missing_columns	available_values_in_missing column
HQ_ADDRESS_COUNTY	79345	0.93	84	6266
MF_INVOLVED_OTHER	71434	0.83	741	14177
MF_TYPE_OTHER	82557	0.96	328	3054
DECADE_INSTALLED	77247	0.90	10	8364
MANUFACTURE_OTHER	82683	0.97	217	2928
LOT_ATTRIBUTES	75017	0.88	738	10594
FITTING_MATERIAL_OTHER	84411	0.99	58	1200
FIRST_PIPE_MATERIAL_OTHER	82845	0.97	45	2766
FIRST_PIPE_PLASTIC_OTHER	85557	1.00	31	54
SECOND_PIPE_MATERIAL_OTHER	85274	1.00	56	337
SECOND_PIPE_PLASTIC_OTHER	85568	1.00	24	43
LEAK_CAUSE_OTHER	79267	0.93	822	6344
NF_THERMAL_EXP_TEXT	72766	0.85	2	12845
EX_DAMAGE_OCCURRED_TEXT	83503	0.98	2	2108
MAT_WELDS_LEAK_CAUSE_TEXT	73060	0.85	2	12551

Benefit:

- Made it harder for me to drop all 15 columns missing >70% values
- Encouraged me to think how we can extract useful information.
- As a result: Replaced **18,789** missing values in 9 columns.

Exploratory Analysis:

Most Used Pipe Size

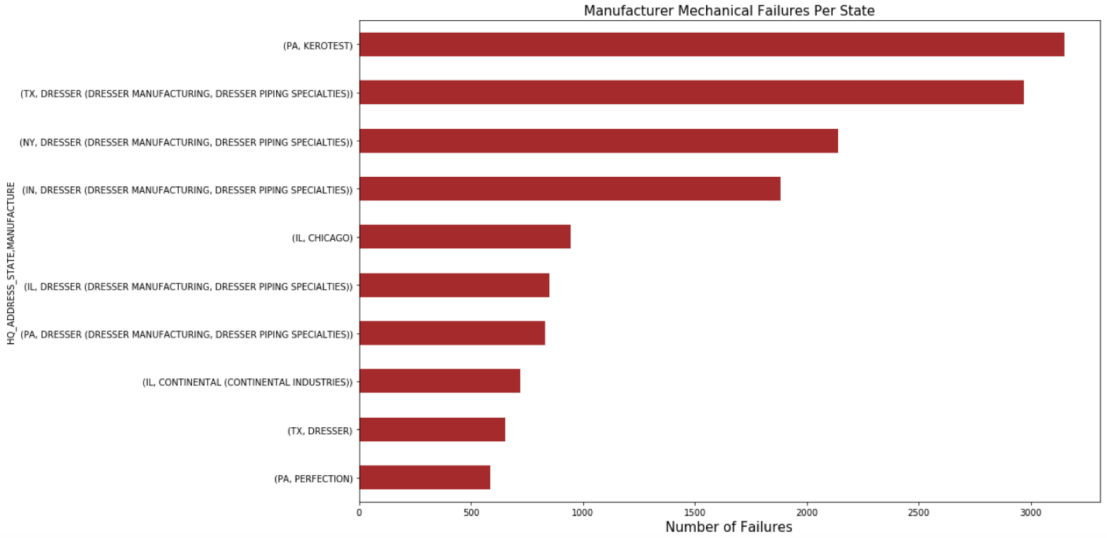


Pipe size 0.75 showed most usage among pipeline operators.

Top 10 Manufacturers by State where leak occurred

HQ_ADDRESS_STATE	MANUFACTURE	
PA	KEROTEST	3152
TX	DRESSER (DRESSER MANUFACTURING, DRESSER PIPING SPECIALTIES)	2972
NY	DRESSER (DRESSER MANUFACTURING, DRESSER PIPING SPECIALTIES)	2138
IN	DRESSER (DRESSER MANUFACTURING, DRESSER PIPING SPECIALTIES)	1881
IL	CHICAGO	944

Name: MANUFACTURE, dtype: int64



PA, Kerotest had maximum leaks

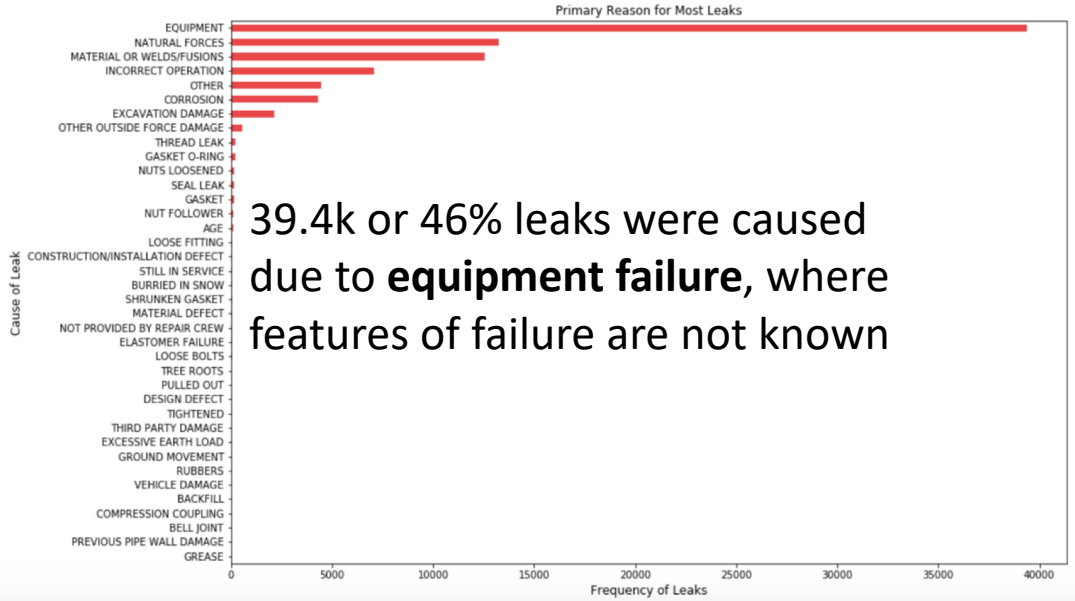
...contd.

Exploratory Analysis:

Top 15 reasons for the leak

****Top 15 reasons for leak****

	Reason_Count	% Reason_Count
EQUIPMENT	39370	45.987081
NATURAL FORCES	13230	15.453622
MATERIAL OR WELDS/FUSIONS	12551	14.660499
INCORRECT OPERATION	7070	8.258285
OTHER	4464	5.214283
CORROSION	4330	5.057761
EXCAVATION DAMAGE	2123	2.479822
OTHER OUTSIDE FORCE DAMAGE	575	0.671643
THREAD LEAK	225	0.262817
GASKET O-RING	223	0.260481
NUTS LOOSENED	180	0.210253
SEAL LEAK	163	0.190396
GASKET	144	0.168203
NUT FOLLOWER	133	0.155354
AGE	90	0.105127



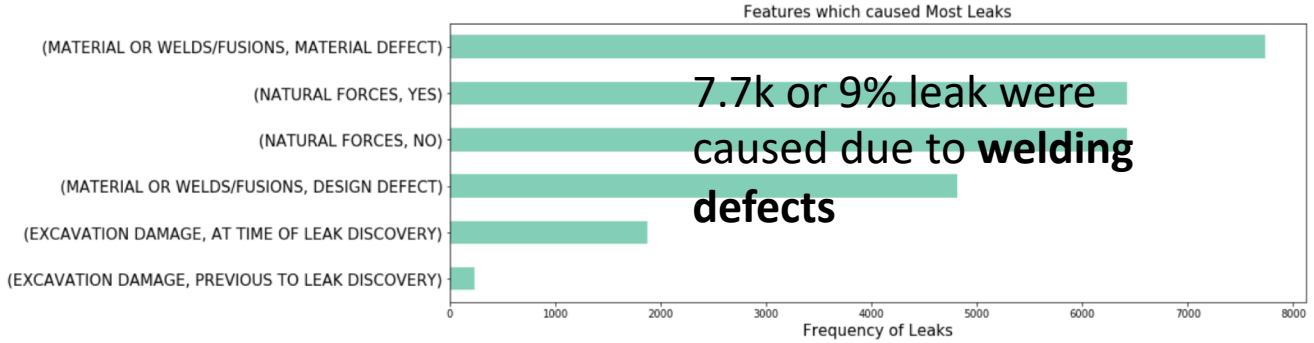
39.4k or 46% leaks were caused due to **equipment failure**, where features of failure are not known

Features which caused leak in the pipeline

****Features which caused leak****

		Feature_Count
LEAK_CAUSE_TEXT	ADDITIONAL_LEAK_FEATURES	
MATERIAL OR WELDS/FUSIONS	MATERIAL DEFECT	7734
NATURAL FORCES	YES	6424
	NO	6421
MATERIAL OR WELDS/FUSIONS	DESIGN DEFECT	4817
EXCAVATION DAMAGE	AT TIME OF LEAK DISCOVERY	1872
	PREVIOUS TO LEAK DISCOVERY	236

		%_Feature_Count
LEAK_CAUSE_TEXT	ADDITIONAL_LEAK_FEATURES	
MATERIAL OR WELDS/FUSIONS	MATERIAL DEFECT	9.033886
NATURAL FORCES	YES	7.503709
	NO	7.500204
MATERIAL OR WELDS/FUSIONS	DESIGN DEFECT	5.626613
EXCAVATION DAMAGE	AT TIME OF LEAK DISCOVERY	2.186635
	PREVIOUS TO LEAK DISCOVERY	0.275666



7.7k or 9% leak were caused due to **welding defects**

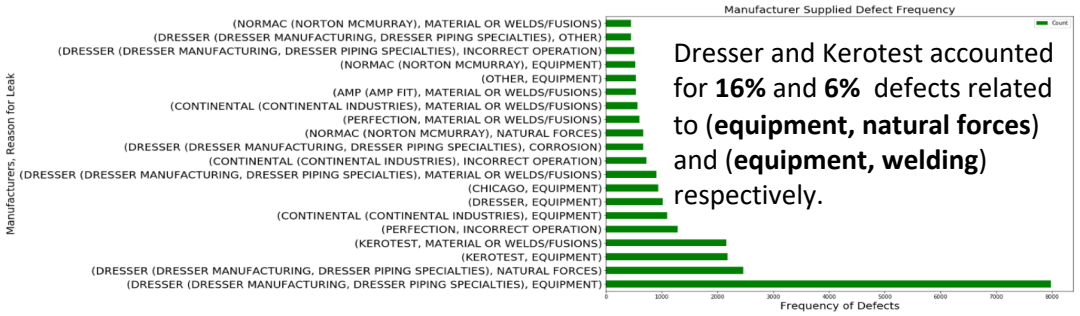
...contd.

Exploratory Analysis:

Known Manufacturer Defects

Manufacturer Supplied Defect Frequency:

MANUFACTURE	LEAK_CAUSE_TEXT	Count
DRESSER (DRESSER MANUFACTURING, DRESSER PIPING ...	EQUIPMENT	7985
	NATURAL FORCES	2461
KEROTEST	EQUIPMENT	2177
	MATERIAL OR WELDS/FUSIONS	2162
PERFECTION	INCORRECT OPERATION	1285

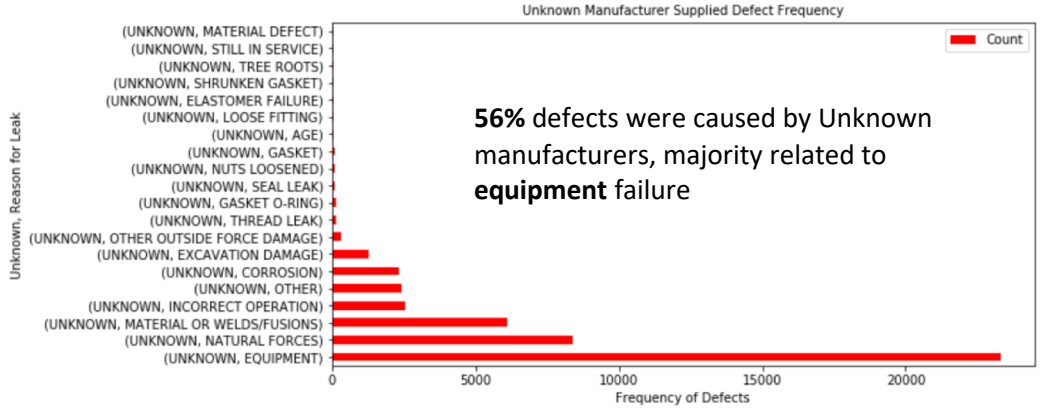


Dresser and Kerotest accounted for **16%** and **6%** defects related to (**equipment, natural forces**) and (**equipment, welding**) respectively.

Unknown Manufacturer Defects

Unknown Manufacturer Supplied Defect Frequency:

MANUFACTURE	LEAK_CAUSE_TEXT	Count
UNKNOWN	EQUIPMENT	23328
	NATURAL FORCES	8408
	MATERIAL OR WELDS/FUSIONS	6108
	INCORRECT OPERATION	2537
	OTHER	2434

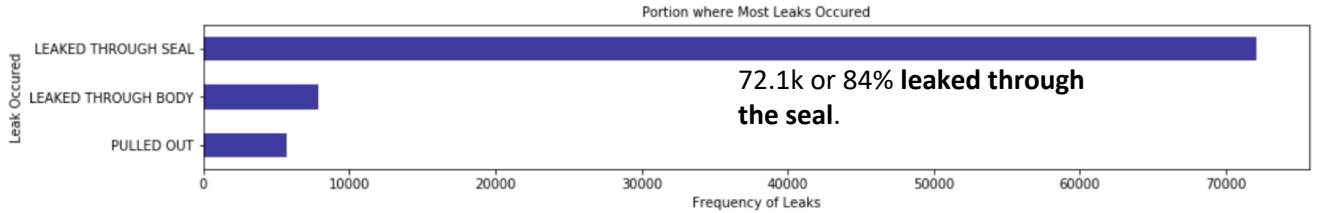


56% defects were caused by Unknown manufacturers, majority related to **equipment** failure

Portion where most leaks occurred

****Portion where leak occured****

	Occurred_Count	% Occurred_Count
LEAKED THROUGH SEAL	72062	84.173763
LEAKED THROUGH BODY	7867	9.189240
PULLED OUT	5682	6.636998

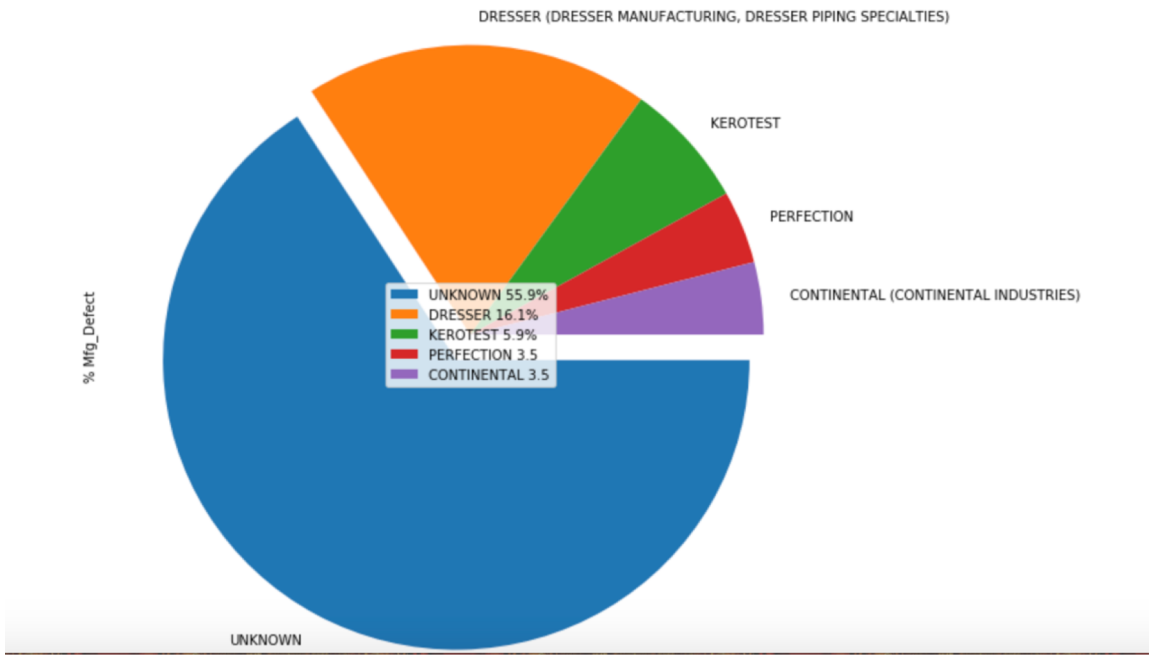


72.1k or 84% **leaked through the seal**.

...contd.

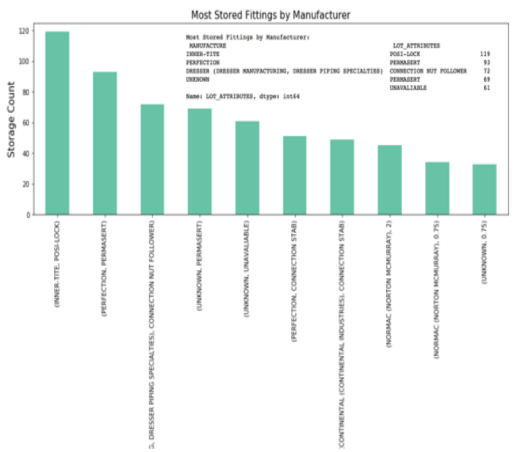
Exploratory Analysis:

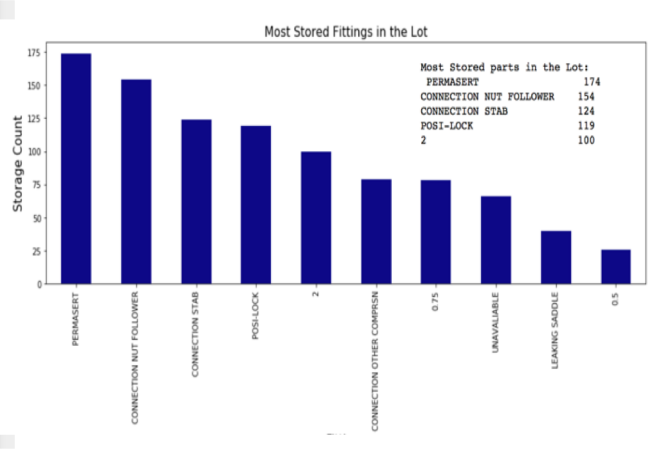
Percentage of Manufacturer Defects



Highest percentage is due to Unknown

Most Stored Fittings by Manufacturer and By Lot





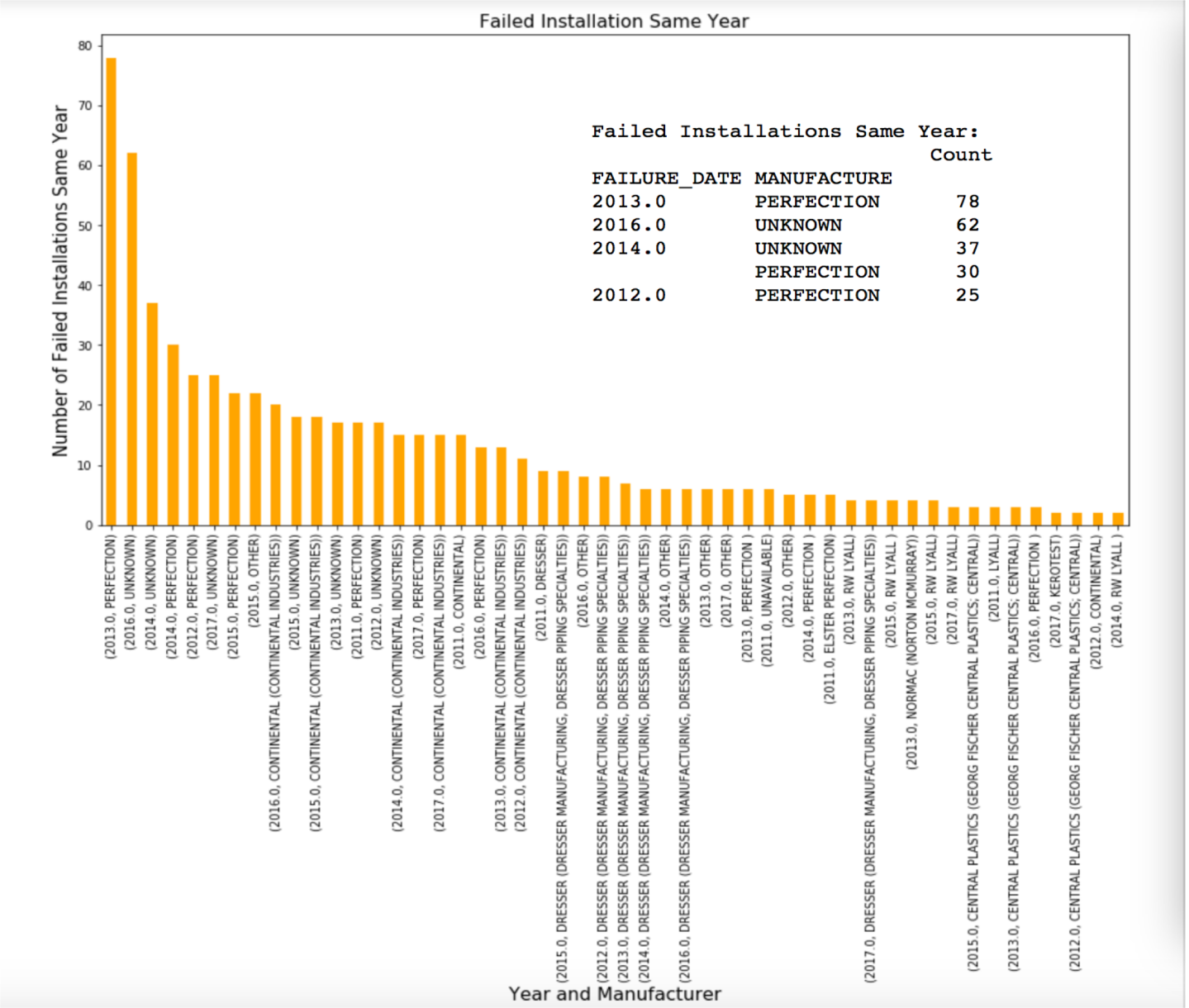
Inter-tite carried maximum POSI-LOCK fittings by manufacturer individually.

Permasert had most stored fitting in the lot supplied by Dresser and Perfection in a lot.

Exploratory Analysis:

Failed Installation Same Year

- Perfection had maximum number of failed installations in 2013.
- Dresser and Continental and Perfection had multiple failed installations between 2011 to 2017.



To be continued…
Data Conversion
ML selection/Evaluation